Chromosome conformation by crosslinking

Polymer physics matters

Jörg Langowski

Division Biophysics of Macromolecules; German Cancer Research Center (DKFZ); Heidelberg, Germany

Correspondence to: Jörg Langowski; Email: jl@dkfz.de

The genetic information is encoded on double-stranded DNA, a long linear polymer chain. Among the central themes of *Nucleus* will be the advancement of our understanding of how those chains are folded so that they fit into the cell nucleus, and at the same time their information can be read off efficiently. In fact, a quantitative description of the structure of the folded genome is one of the most challenging problems in structural biology, and poses a much more formidable problem than—for instance—the folding of a protein. There are three main reasons for this: first, the genomic DNA is by orders of magnitude the largest biomolecule in the cell; second, it cannot be defined by a single spatial structure because of its flexibility; and third, even if the 'fold' of the genome were more or less defined in any one cell, variations among individual cells may be very large, for the same reasons of flexibility.

Chromosome Conformation Capture

Recently, the method of chromosome conformation capture (3C) and its variants have provided a new, very powerful tool for looking at interactions among different regions of the chromosomes in the cell. The principle is simple and elegant: use a crosslinking agent such as formaldehyde and covalently link those parts of the chromatin that are close in space. Then, digest the chromatin with restriction enzymes, ligate the ends of crosslinked fragments using a DNA concentration low enough to favor intramolecular ligation, reverse the crosslinks, deproteinize and then PCR amplify

the DNA with suitable primers to identify pairs of genomic regions that had been linked together. In their original work,[3] Job Dekker and colleagues measured the interaction between pairs of a small number of specific loci in yeast. Over the last years, the technique has seen impressive development, and a number of improvements have been added to facilitate its application to an ever-increasing number of genetic loci. 3C and its variants have been used very widely since the original work came out; as of this writing, the paper has been cited 290 times.

Although at the time the authors proposed a structural model for yeast chromosome III, that model had to be taken with a large grain of salt. Very probably, the mechanical flexibility of DNA and the chromatin fiber randomizes any particular chromatin fold to such an extent that the 'average' structure will bear very little resemblance to the actual fold of the chromosome in a particular instant. This is seen, for instance, in the large motions exhibited by fluorescently labeled parts of the genome during live cell imaging.[4-6] Thus, defining one single average structure of a large flexible biopolymer makes about as much sense as defining the average structure of a football player during a game. The work discussed here takes this randomness very clearly into account and still arrives at very profound conclusions about genome organization.

Large-Scale 3C: Identifying Crosslinked Regions by High Throughput Sequencing

It is an intriguing idea to use 3C or a related technology for analyzing the folding of a

complete genome. While higher-through-put variants of 3C have been developed in the last years, only very recently has the coupling of genome crosslinking with massively parallel sequencing enabled tackling problems of that size. I would like to discuss two recent papers that applied the crosslinking/sequencing strategy to the human and to the yeast genome.

Crosslinking Human Cell Lines

The strategy chosen by Lieberman-Aiden et al.[1] is called Hi-C. After formaldehyde crosslinking, they cut the genomic DNA with a restriction enzyme that leaves over-hanging ends and fill them in with biotiny-lated nucleotides. Blunt-end ligation created chimeric DNAs in which the interacting parts of the genome are covalently linked together and which carry a biotin at the site of ligation. This enabled them, after shear-ing the DNA to a convenient size, to purify the chimeric fragments using streptavidin-coated paramagnetic beads, and to analyze the fragments by Illumina sequencing.

All in all, more than eight million sequence reads were aligned to the human genome, each of them corresponding to a pair of interacting sequences. The num-ber of interactions between each pair of 1 Mb genome segments was represented in a two-dimensional matrix, and the assumption made that the higher this number, the stronger was the interaction between the corresponding two segments.

This 'heatmap' immediately shows some interesting features. First of all, the contact probability between the loci in one chromosome was always larger than the contact probability between different chromosomes. This proves directly the existence of chromosome territories, i.e., that most of the mass of a given interphase chromosome occupies a volume that is not extensively invaded by other chromo-somes. Also, the small, gene-rich chromo-somes, which FISH studies locate rather in the center of the nucleus, were found to interact preferentially with each other.

'Block Matrix' Structure of Genomic Interactions

Within chromosomes, certain regions showed a higher and others a lower

interaction frequency relative to the aver-age interaction probability for loci at that distance. While this was not surprising per se, the interesting result was that these regions of increased and decreased inter-action frequency are not distributed ran-domly; they come in rather large blocks comprising about one-tenth of the whole chromosome. First, this shows the exis-tence of chromosomal subcompartments, which had been postulated some years ago from FISH experiments and simula-tion studies.[7-9] Second, the whole genome seems to be divided into two compart-ments, one of which contains more genes, is more highly transcribed and more accessible to DNase than the other com-partment. Finally, a powerful statistical analysis revealed that if two chromosome interact, their open regions preferentially interact with one another, and the same holds for the compact regions. The anal-ogy with the classical cytological entities of euchromatin vs. heterochromatin obvi-ously presents itself.

Fractal Globule Folding

What does the ever-descending subcom-partmentalization imply for the overall folding of the chromosomes in the cell nucleus? One way of characterizing the random folding of a polymer chain is to analyze the variation of the interaction probability $I(s)$ between two loci at a dis-tance s. Lieberman-Aiden's data showed a power-law behavior, with $I(s)$ decreasing as s^{-1} for genomic distances up to the size of the subcompartments. Again, this implies a particular folding pattern even within the subcompartment, since for a flexible polymer simply stuffed randomly into a sphere (a so-called 'equilibrium globule') their Monte-Carlo simulations showed that $I(s)$ will be proportional to $s^{-3/2}$, the same as the well-known behavior of an unconstrained chain.[10]

This means that on the scale of several megabases, chromatin cannot be simply folded randomly, but consists of very small 'globular' regions which condense to larger globules, which then form even larger globules et seq., much as in the Jonathan Swift verse: "So nat'ralists observe, a flea/ Hath smaller fleas that on him prey,/And these have smaller fleas that bite 'em,/And

so proceed ad infinitum." Chromosome subcompartmentalization does not really "proceed ad infinitum," it only extends from the level of whole chromosomes down to a scale of some hundred kilobases, i.e., of the order of a thousand nucleosomes. Still, the resulting structure is self-similar (i.e., fractal) over at least two orders of magnitude. In such a 'fractal globule' the interaction probability decreases with the inverse of the genomic distance, just as the experiments showed.

The fractal globule structure has interesting implications for chromatin unfolding: the structure can freely expand without entanglements, which is much more difficult in a randomly folded chain because it is highly knotted. The seeming ease with which the interphase chromo-somes transition into their familiar, non-tangled early prophase configurations is compatible with this notion.

Crosslinking Yeast

It is interesting to compare Lieberman-Aiden's data in human chromosomes with another, very similar data set obtained practically at the same time by Rodley et al. in yeast.[2] The yeast genome being much smaller that the human one, the crosslinking/ligation/sequencing strategy can do without the biotin/streptavidin puri-fication step. Cutting the formaldehyde-crosslinked genome with an appropriate restriction enzyme, then intramolecularly ligating fragment ends, shearing the liga-tion mix and Illumina-sequencing the generated fragments showed a regenerated restriction site (indicative of a crosslink) in about 2% of the sequences. The rest was used for genome assembly.

Similar to the human genome data, Rodley and colleagues generated interac-tion maps within and between the chro-mosomes, the mitochondrial DNA and the 2 μ plasmid. While globally the num-ber of unique interactions increased lin-early with the length of each chromosome, not all possible chromosome pairings were observed. Loops within chromosomes could be clearly identified, and a multi-looped structure was postulated for yeast chromosomes, reminiscent of the con-cepts in the multi-loop subcompartment model by Münkel et al.[7] or the random-

loop model by Bohn et al.[11] An interesting difference was found for essential vs. non-essential genes: the former were shown to be interacting much less with distant parts of the genome. The authors postulated that such genes might be segregated into separate subdomains for better accessibility.

Metabolic conditions were observed to influence intragenomic interactions: changing the carbon source from glucose to glycerol/lactate and galactose clearly changed the frequency of interactions of selected loci on the genome, 2 μ plasmid and mitochondria. The paper also showed that the 2 μ plasmid is folded such as to maximize interactions between inverted repeats and expose the stability locus, maybe to facilitate clustering. Finally, specific interactions between the yeast chromosomes and the mitochondrial genome were demonstrated, which may be specific and used in regulation.

In general, the crosslinking data in yeast might help us to understand local, specific interactions better than data collected on the much larger scale of a higher eukaryote. In that aspect, the two approaches very nicely complement each other.

Synopsis: Models are Needed

Of course, interaction probability cannot be directly related to spatial distance. While there is a unique relation between $I(s)$ and s for a Gaussian random flight polymer chain, there are many circumstances that can modify this behavior or even render this relationship completely invalid. A global correlation between $I(s)$ and s can be drawn for the human and yeast data set, but since the data is an average over a large number of cells, one cannot be sure whether this ensemble average also reflects the time average, which one would obtain by measuring a single cell over the course of its cell cycle. Second, even if the ensemble and time averages are the same, the exact relation between $I(s)$ and s is still unknown. Third, the crosslinking probability may be greatly influenced by the local folding of chromatin around the interacting sites, and by associated chromatin-binding proteins. All these effects render a direct interpretation of crosslinking data in terms of intragenomic distances rather uncertain. However, the *scaling* of the interaction probability with genomic distance remains an important parameter characterizing the overall folding. Such scaling concepts have been an important part of polymer theory since the pioneering work of de Gennes[12] and can provide deep insight into the random structure of a macromolecule.

To go beyond scaling laws, a direct measurement of $I(s)$ as a complement to the crosslinking data and as input into a more detailed model would be very desirable. This can be done (and has been done for small numbers of interacting loci) by imaging, using FISH or related techniques. While such measurements are very time-consuming compared to the massively parallel Hi-C, GCC or the 3-, 4- and 5-C procedures, they may be used for calibration of the crosslinking techniques and as input to folding models such as the one proposed in the Lieberman-Aiden article, or others that regard the chromatin fiber as a flexible polymer chain.[8,11,13,14] The number of conformations accessible to a folded chromatin chain in the cell is huge. To compare them with the experimental data and pick out the most probable ones, both advanced computer modeling techniques and large amounts of experimental data are needed. The two papers discussed here provide the latter and thus constitute an important step toward a comprehensive model of genome folding.

References

1. Lieberman-Aiden E, van Berkum NL, Williams L, Imakaev M, Ragoczy T, Telling A, et al. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. Science 2009; 326:289-93.
2. Rodley CD, Bertels F, Jones B, O'Sullivan JM. Global identification of yeast chromosome interactions using Genome conformation capture. Fungal Genet Biol 2009; 46:879-86.
3. Dekker J, Rippe K, Dekker M, Kleckner N. Capturing chromosome conformation. Science 2002; 295:1306-11.
4. Marshall WF, Straight A, Marko JF, Swedlow J, Dernburg A, Belmont A, et al. Interphase chromosomes undergo constrained diffusional motion in living cells. Curr Biol 1997; 7:930-9.
5. Heun P, Laroche T, Shimada K, Furrer P, Gasser SM. Chromosome dynamics in the yeast interphase nucleus. Science 2001; 294:2181-6.
6. O'Brien TP, Bult CJ, Cremer C, Grunze M, Knowles BB, Langowski J, et al. Genome function and nuclear architecture: from gene expression to nanoscience. Genome Res 2003; 13:1029-41.
7. Münkel C, Eils R, Dietzel S, Zink D, Mehring C, Wedemann G, et al. Compartmentalization of interphase chromosomes observed in simulation and experiment. J Mol Biol 1999; 285:1053-65.
8. Mateos-Langerak J, Bohn M, de Leeuw W, Giromus O, Manders EM, Verschure PJ, et al. Spatially confined folding of chromatin in the interphase nucleus. Proc Natl Acad Sci USA 2009; 106:3812-7.
9. Zink D, Cremer T, Saffrich R, Fischer R, Trendelenburg MF, Ansorge W, et al. Structure and dynamics of human interphase chromosome territories in vivo. Hum Genet 1998; 102:241-51.
10. Jacobson H, Stockmayer WH. Intramolecular reaction in polycondensations I. The theory of linear systems. J Chem Phys 1950; 18:1600-6.
11. Bohn M, Heermann DW, van Driel R. Random loop model for long polymers. Phys Rev E Stat Nonlin Soft Matter Phys 2007; 76:051805.
12. de Gennes PG. Scaling concepts in polymer physics. Cornell University Press, Ithaca 1979.
13. Münkel C, Langowski J. Chromosome structure described by a polymer model. Physical Review E 1998; 57:5888-96.
14. Rosa A, Everaers R. Structure and dynamics of interphase chromosomes. PLoS Comput Biol 2008; 4:1000153.